The NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR)

Evaluation Specification Document

Version: mm08_evalplan_v1.1 *Dated:* April 4 2008

1 Introduction

NIST has been conducting formal evaluations of machine translation (MT) technology since 2002, and while the evaluations have been successful, there is still a need for a better understanding of exactly how useful the state-of-the-art technology is, and how to best interpret the scores reported during evaluation.

This need exists primarily due to the shortcomings with the current methods employed for the evaluation of Machine Translation technology:

1) Automatic metrics have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies.

2) Automatic metrics have not demonstrated that they are as meaningful in target languages other than English.

3) Human assessments are expensive, slow, subjective, and are difficult to standardize[1]. Furthermore they only pertain to the translations evaluated, and are of no use even to updated translations from the same system.

4) Both automatic metrics and human assessments need more insights into what properties of the translation should be evaluated, as well as insights into how to evaluate those properties.

5) Some MT technology approaches evaluated incorporate algorithms that optimize scores on MT metric(s). These optimizations fail in the same respects that the metrics fail.

These problems, and the need to overcome them through the development of improved automatic (and even semi-automatic) metrics, have been a constant point of discussion at past NIST MT evaluations. Without more appropriate metrics to address these shortcomings, the impact of formative and summative MT technology evaluations will remain limited.

To date, the evaluation of MT metrics has been a secondary issue, lagging behind the effort put towards the evaluation of MT technology. Developers of metrics conceive, implement and test their ideas independently. Papers are written and published and some metrics have been adopted for use in large scale evaluations. Unfortunately, there has only been limited focus on the development and assessment [2]of the metrics themselves.

NIST, therefore, will be conducting a new MT evaluation series focused entirely on MT metrology, where innovative MT metrics are being evaluated rather than MT technology.

NIST will install the software for participants' metrics and run them locally on a set of carefully selected system translations.

[1] Human assessments can be tailored to a specific application of interest. For the MetricsMATR evaluation, NIST used the implementation designed for NIST Open MT-2008. See http://mt.nist.gov/MT1/login.php (use "guest" for both the username and the password) for an overview of the TAP-ET application used.

[2] An evaluation of MT metrics is scheduled as part of an ACL 2008 workshop, "Third Workshop on Statistical Machine Translation, ACL08-SMT". See http://www.ling.ohio-state.edu/acl08/cfw.html.

The first such evaluation cycle will conclude with a full-day workshop at AMTA 2008, "MetricsMATR: The Metrics for Machine Translation Challenge". The goal of the workshop is to inform other MT technology evaluation campaigns and conferences with regard to improved metrology. The first MetricsMATR challenge will establish the metric evaluation infrastructure that encourages the development of (preferably automatic) metrics addressing the shortcomings identified above. The MetricsMATR challenge will include an evaluation of all the submitted metrics that compares performance on a substantial set of MT outputs (initially, with English as the target language).

MetricsMATR will not exclude modifications and extensions of existing metrics. However, there will be a strong emphasis on the development of clearly innovative, even revolutionary, metrics that have the potential to once more initiate a substantial paradigm shift in the field of MT metrology, much as the introduction of BLEU[3] did in 2001.

The MetricsMATR challenge is designed to appeal to a wide and varied audience including researchers of MT technology and metrology, acquisition programs such as SEQUOYAH, and commercial vendors. We will welcome submissions from a wide range of disciplines including computer science, statistics, mathematics, linguistics, and psychology. NIST encourages submissions from participants not currently active in the field of MT.

2 Data

Two separate data sets are being developed to support the MetricsMATR evaluation. The first is a development data set that will be made available to developers of metrics preparing to participate in the 2008 evaluation. The second is the evaluation data set that will be sequestered by NIST.[4]

2.1 Development Data

The development data set (dev-set) will be distributed to metric developers upon registering for participation in the evaluation.[5] (See section 5 for the target availability date of the dev-set.) The development data will consist of:

- Several versions of system translations
- Up to four independently created reference translations
- Segment level human assessments of *adequacy*
    - Document and system level scores will be created as described in section 4.
- Segment level human judgments of *preferences*
- (Source translations will be available on request, may require a separate license agreement)

In addition to adequacy and preference, there are many other forms of human assessments including fluency, concept transfer, and edit-distance. While research in the area of various human assessments is ongoing, resource restrictions limit the MetricsMATR evaluation to *adequacy* and *preference* assessments.

The dev-set comes from the NIST Open MT-06[6] evaluation and from DARPA TRANSTAC training dialogs. The documents from the MT-06 evaluation were selected by examining the document level BLEU scores across several

[3] Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. (2002), "BLEU: a method for automatic evaluation of machine translation", in ACL-2002: 40th *Annual meeting of the Association for Computational Linguistics*, pp. 311-318.

[4] We envision that the evaluation data set will be reused and expanded in future MetricsMATR evaluations. Some of the evaluation data is owned by other HLT programs and NIST does not have permission to distribute the system translations.

[5] To register your participation fill out the forms located at http://www.nist.gov/speech/tests/metricsmatr/2008/doc or send e-mail to mt_poc@nist.gov for more information.

[6] The 2006 NIST Open MT evaluation is documented here: http://www.nist.gov/speech/tests/mt/2006/doc. The evaluation test set is owned by the Linguistic Data Consortium, see http://www.ldc.upenn.edu.

systems. NIST hand-selected each document to provide varying levels of performance, as determined by the MT-06 official evaluation metric, BLEU. See Table 1 for dev-test corpus information. For the TRANSTAC dialogs, the offline training data was readily available and was included to provide some sampling of an alternative data style.

Table 1: MetricsMATR Dev-Set Statistics

Source of Data	MT-06	TRANSTAC
Genre	newswire	training dialogs
Number of documents/scenario	25	1*
Total number of segments	249	17
Source Language	Arabic	Iraqi Arabic
Number of system translations	8	5

* The single document for the TRANSTAC training dialog contains 17 segment of unrelated text.

2.1.1 System Translations

Translations from multiple systems are included in the dev-set. Five systems are included from the January 2007 TRANSTAC offline evaluation. Eight systems were selected from the MT-06 evaluation, selected to cover a variety of MT algorithmic approaches (statistical MT, rule-based, hybrid …). All system names will be anonymous.

2.1.2 Reference Translations

MT-06 reference translations are provided. The Linguistic Data Consortium supplied four independently created translations for each document. Each translation agency was given the same set of translation guidelines[7].

References for the TRANSTAC dialog data were created by Appen using specific guidelines for transcription and translation.

2.1.3 Human Assessments of Adequacy

Assessments of adequacy were performed using the NIST TAP-ET application[8]. Each translation segment was assessed by two judges according to the guidelines described in the application documentation. After completely assessing the entire dev-set, the judges reviewed their individual assessments together and settled on a single final score.

2.1.4 Human Judgments of Preferences

Assessments of preferences were performed using the NIST TAP-ET application. Each translation segment was compared to each and every other corresponding translation by two judges, according to the guidelines described in the application documentation. After completely assessing the entire dev-set for preferences, the judges reviewed their individual preferences together and settled on a single final preference.

2.1.5 Source Transcriptions

Source transcriptions exist and can be made available if required by the developed metric. Contact NIST mt_poc@nist.gov to discuss the implications of signing the required license agreement.

2.2 Evaluation Data

The evaluation data set (eval-set) will not be distributed to participants. To the extent possible, NIST will leverage data resources with the most recent MT-08, GALE, and TRANSTAC evaluations. Data from other sources *may* be included. The eval-set will be similar to, but more expansive than the dev-set and will include different systems and data translated from different source languages that were not represented in the dev-set allowing for analysis

[7] http://projects.ldc.upenn.edu/translation/MT08
[8] http://www.nist.gov/speech/tools

on data with properties on which the metrics could not have been specifically tuned. To the extent possible the evaluation data will be categorized in ways that might assist in interpretation of the results (ILR levels, genres …).

The same procedures for assessing adequacy and preference judgments are used for both the development and evaluation data sets.

2.3 File Formats

This section describes the file format for both the input files that the metrics will be required to read, and the output files that the metrics should produce.

2.3.1 Metric Input Files (System Translations and Reference Translations)

Input files will be in an XML format that NIST uses for the Open MT evaluations. NIST has defined a set of XML tags that are used to format MT source, translation, and reference files for evaluation. Each set of translations for a single system will be identified in separate files. See Appendix A for detailed file format information.

2.3.2 Metric Output Files

Analysis of the submitted metrics will take place on various levels. MetricsMATR prefers metrics be designed to output system, document, and segment level scores, but it may be the case that a metric is not designed to do so. In such cases, please alert NIST before the evaluation so we can prepare accordingly.

Metric developers must output scores in the format described below. This will allow for plug-in comparisons for the various correlation tests, and it will significantly reduce the possibility of human-introduced errors in a reformatting process.

One running of the software on a single translation file should produce at least (3) files:

1. <System Name>-sys.scr *# System level scores*
2. <System Name>-doc.scr *# Document level scores*
3. <System Name>-seg.scr *# Segment level scores*

Contents of these three files are described below.

2.3.2.1 System Scores

The evaluated metric should have the capability for assigning a **single overall "score" for a system**. To assist in analysis, we are requiring the metric to output a system level score file "**<System Name>-sys.scr**" for each input file evaluated. The output should be a single tab separated record:

```
<TEST_ID>   <SYSTEM_ID>   <SYSTEM LEVEL SCORE>   <OPTIONAL>
```

Where:

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
SYSTEM LEVEL SCORE is the overall system level score.

Followed by optionally included items each separated by a tab (confidence scores, statistics ...).

2.3.2.2 Document Scores

The evaluated metric should have the capability for assigning a score to each **document** translated by the system. Note for some data types (e.g., transcripts of dialogs), document is the term we will use to refer to a single grouped exchange, scenario, or discussion. To assist in analysis, we are requiring the metric to output a document

level score file "**<System Name>-doc.scr**" for each input file evaluated. The output should be a single tab separated record for each document:

```
<TEST_ID>   <SYSTEM_ID>   <DOCUMENT_ID>   <DOCUMENT LEVEL SCORE>   <OPTIONAL>
```

Where:

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
DOCUMENT_ID is the document identified by the **docid** attribute in the translation file.
DOCUMENT LEVEL SCORE is the overall document score.

Followed by optionally included items each separated by a tab (confidence scores, statistics ...).

2.3.2.3 Segment Scores

The evaluated metric should have the capability for assigning a score to each **segment** translated by the system. To assist in analysis, we are requiring the metric to output a system level score file "**<System Name>-seg.scr**" for each input file evaluated. The output should be a single tab separated record for each segment:

```
<TEST_ID> <SYSTEM_ID> <DOCUMENT_ID> <SEGMENT_ID> <SEGMENT SCORE> <OPTIONAL>
```

Where:

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
DOCUMENT_ID is the document identified by the **docid** attribute in the translation file.
SEGMENT_ID is the segment identified by the **id** attribute of the **seg** tag.
SEGMENT SCORE is the score for the particular segment.

Followed by optionally included items each separated by a tab (confidence scores, statistics ...).

3 Evaluation Tracks

Metric developers are required to develop software that implements a scoring algorithm which assesses machine translation quality. The scoring software is to be packaged and submitted to NIST for evaluation. The submitted package should identify system requirements, minimum versions of installed tools, and a short description of interpretation of the scores (error, accuracy …). NIST will install and use each participant's code to evaluate output from a variety of machine translation systems. Each submitted metric will be graded by how well the properties of the scoring output correlate to carefully annotated human assessments of adequacy and comparative preferences between translations.

Participants are encouraged to be adventurous and creative in their metric development and to avoid being overly influenced by techniques that have already been attempted.

See section 7 for some practical limitations and guidelines for metric development.

Metrics-MATR will analyze the submitted metrics in two tracks, differing by the number of reference translations available.

3.1 Single Reference track

There is a cost associated with creating reference translations for use in evaluation. If metrics were determined to be great predictors of MT quality based on only one manual reference translation instead of multiples,

evaluation data sets could grow in size. This in turn would provide better grounds for determining statistical differences.

For the "Single Reference Track", NIST will analyze the metric performance when limiting all of the evaluation data to one pre-selected reference translation per test segment.

3.2 Multiple References track

Most will agree that often there is not a single "best" or "perfect" translation of every given source sentence. For some language pairs, such a perfect translation is not possible. There are other issues as well, such as multiple acceptable ways to handle idioms, name variants, and synonymy.

Previous experiments have found metrics that make use of more than one independently created reference translation tend to asymptote for metric stability as measured against human judgments of quality, around the use of four references. For much of the data, four reference translations will be available.

NIST will analyze submitted metric performance separately for metrics that are designed to use more than one reference. The reference translations will be ranked identifying which is to be used for the single reference track.

4 Evaluating the Metrics

4.1 Correlation with Human Judgments

For this evaluation, the reference is not the true translation, but rather the agreed-upon subjective grade of adequacy (or preference) assigned to each translation. The method for grading the submitted metrics is defined to be how well the produced scores correlate with these grades, and how well they can provide insight as to the quality of the MT translations.

4.1.1 Segments

The basic assessments were performed at the isolated segment level although segments were presented in document order. Each segment received two independent judgments, and those two judgments were then adjudicated into one score.

4.1.2 Documents

NIST will create "document level" assessment scores by combining the set of segment scores. NIST will experiment with a few methods of combination including, but not limited to:

1. the mean of the segment scores
2. a weighted-average score (taking into account segment length)

The preferred score for document-level evaluation will be the second weighted method.

4.1.3 Systems

NIST will create "system level" assessment scores by combining the set of segment or document scores. NIST will experiment with a few methods of combination including, but not limited to:

1. the mean of the segment scores
2. a weighted average score (taking into account segment length)
3. the mean of the assigned document scores

The preferred score for system-level evaluation will be the second weighted method.

4.2 Correlation Measures

NIST plans to involve our statistical division to assist in our analysis and measuring of correlations. As a starting point, we will calculate the correlation coefficients for:

- Pearson's *r*
- Kendall's tau
- Spearman's rho

5 Schedule

The following table outlines the important dates for this evaluation cycle.

Date	Milestone Description
Mar 31 2008	Development data set released
Aug 01 2008	Deadline for participation commitment
Aug 01 2008 – Sep 05 2008	Metric submission period
Sep 05 2008	Deadline for metric software to be installed and operational at NIST
Sep 30 2008	Paper describing submitted metric(s) due at NIST
Oct 25 2008	AMTA 2008 workshop: MetricsMATR

6 Workshop

The report-out session for this evaluation will be an AMTA workshop: "MetricsMATR, The Metrics for Machine Translation Challenge." This will be an open workshop where NIST provides an evaluation overview discussing the metrics submitted. A report will be given that reviews the correlations as described in section 4.

All metric developers participating in MetricsMATR08 must submit a paper describing the submitted metric(s) to NIST by September 30, 2008. Metric developers will be given the opportunity to present their metrics and discuss results achieved using the development data set.

There will be panel discussions and plenty of time dedicated to how best to continue this evaluation series, moving forward with automatic and semi-automatic metrics that overcome the difficulties identified in the Introduction.

7 Metric Properties

There are many desirable properties for metrics employed for evaluation. In this section, we discuss a few general guidelines that should be considered during metric development. This list is known to be incomplete and other evaluations may have contradictory thoughts. Section 7.1 describes properties required for practical implementation and testing of metrics, and section 7.2 describes the characteristics or capabilities of metrics that are so clearly missing.

7.1 Practical Implementation Properties

The following set of properties describes practical limitations that all automatic metrics will possess in order to be deemed useful for evaluation purposes.

7.1.1 Automaticity

Metrics that are "automatic", that is, metrics that do not require human intervention outside the creation of the reference translations, are useful to evaluate systems over large test sets. Large test sets lead to greater power of

statistical tests and allow for evaluation over greater populations. Automatic metrics can also be used in training by certain MT technology approaches.

7.1.2 Repeatability (Reliability)

It is extremely desirable that metrics produce the exact same score each time they are used to evaluate the same set of data.

7.1.3 Portability

Metric software should be universally usable. Metrics should not require support of antiquated software, or unusual operating systems. It is expected that a knowledgeable system administrator will be able to install and compile all components of the developed metric within a half day's work.

The metric might make use of tools on the internet. They should be failsafe in case the internet is unavailable.

For MetricsMATR, the software will need to run on at least one of the following operating systems:

1. Windows XP
2. MAC OS X
3. Linux CENT OS 5 (or newer)

7.1.4 Speed

Metric software should be relatively quick to run. If a metric requires more than 5 hours to score the complete dev-set, the developer should contact NIST to discuss other options before the evaluation.

7.1.5 Limited Annotation of Reference Data

The evaluation infrastructure will include up to 4 independently created reference translations[7] for each translated segment. Reference translations are created following standard translation guidelines and do not include additional mark-up for items such as proper names and alternations. Some algorithms may require additional mark-up, but in the implementation of the MetricsMATR's evaluation, the references are not released.

We suggest that in cases where the submitted metrics might benefit from additional mark-up of the references, the possibility for success be demonstrated by an automatic mark-up process in this first year. If promise is shown, NIST may invest in updating the references for future evaluations.

7.2 Metrology Objectives

The following properties are strongly sought after behaviors and capabilities that are currently missing from existing automatic MT metrics.

7.2.1 Correlation with Human Assessments of MT Quality

Currently, the slow, tedious, and subjective process of humans comparing the system translation to the reference translation is one of the most accepted ways of determining which systems are better than others. Thus correlations with human assessments are the primary metric to be used in this evaluation.

7.2.2 Ability to Differentiate Between Systems of Varying Quality

To the extent possible, metrics should be able to differentiate quality between two different systems. That is, the reported scores should be fine-grained enough to rank even systems that are fairly close in quality.

7.2.3 Intuitive Interpretation

A complaint levied against current automatic MT metrics is that the reported score is difficult to relate to quality. This makes it difficult to demonstrate how meaningful MT improvements are. To the extent possible, it is desirable that the reported score be directly related to quality and be intuitive even to persons without specific technical background in machine translation.

7.2.4 Applicability to Multiple Target Languages

While the first implementation of MetricsMATR limits the target language to English, metrics that work on a wide variety of target languages will be of most benefit.

7.2.5 Stability against Optimization

In the framework of this evaluation, the system translations that are evaluated were not optimized for the metrics being developed. There is a chance however, that results on this blind evaluation data set may differ from results on translations that were optimized for the particular metric. The goal is to get away from gaming and metric tuning.

Specific questions related to information in this document should be addressed to: mt_poc@nist.gov

MetricsMATR web-site: http://www.nist.gov/speech/tests/metricsmatr

Appendix A: NIST MT XML Data Format

I. Translation File Format

Each translation file contains translations for a single system to be evaluated. The translation file format is defined by the current MT DTD,[9] and will begin with the following three lines (numbered for identification):

```
1. <?xml version="1.0" encoding="UTF-8"?>
2. <!DOCTYPE mteval SYSTEM ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.0.dtd>
3. <mteval>
```

Line 1: XML header, definition statement
Line 2: DTD identifier
Line 3: MTEVAL tag identifies the beginning of a test set.

A translation section will begin with a **<tstset>** tag which contains a set of documents. Each document, defined by the **<doc>** tag, contains a set of segments. Each segment, defined by the **<seg>** tag, contains the translated text. The source, translation, and reference(s) documents will each contain the same number of segments. (although it is possible for a translation segment to be empty).

The **<tstset>** tag has two required attributes "**setid**" and "**srclang**", and one implied attribute "**trglang"**. The **setid** attribute contains the name of the document set that has been translated. This name will match the **setid** of the source file for which system performed the translation, and the **setid** of the reference file which will be used to evaluate the system translations. The **srclang** attribute indicates the language of the source set and the **trglang** attribute indicates the language of the translate set, for Metric MATR this will be set to "**English"**.

The **<doc>** tag has two required attributes "**docid**" and "**genre**", and one implied attribute "**sysid**". The **docid** attribute contains the name identifying the document within the given source set. The genre attribute indicates the type of data for a given document. The **sysid** attribute contains the name of the system that performed the translation.

The **<seg>** tag has an implied attribute called **id**. The **id** attribute contains a number identifying the segment within the given document. Note that the translation segments must appear in the same order as the order they appear in the source and reference file.

```
4. <tstset setid="mm08_set1_v0" srclang="Arabic" trglang="English">
5. <doc docid="document-1" genre="newswire" sysid="mm08_set1_system1">
6. <seg id="1"> TRANSLATED ENGLISH TEXT. </seg>
7. <seg id="2"> TRANSLATED ENGLISH TEXT. </seg>
8. …
9. </doc>
10. <doc docid="document-2" genre="newswire" sysid="mm08_set1_systems">
11. …
12. </doc>
13. …
14. </tstset>
```

[9] The current version of the NIST XML MT DTD may be found at: ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.0.dtd

15. </mteval>

Line 4: TSTSET tag identifies the beginning of a document list
Line 5: DOC tag, one for each document in the TSTSET
Line 6: SEG tag, ordered 1-N, for each sentence like unit in the document
Line 8: Many possible segments per document
Line 9: Closing DOC tag ends translations for the particular document
Line 13: Many possible documents in each TSTSET
Line 14: Closing TSTSET
Line 15: Closing MTEVAL tag

Note there are other possible tags that may be present in the system translation files. Headline tags and paragraph markers are two of many possible examples. See the MT DTD for a complete description of possible tags.

II. Reference File Format

A single reference file will contain all the reference translations available for an identified data set (**setid**). Some of the system translations will have only one reference translation, while others will have 4. The reference file format is defined by the current MT DTD, and will begin with the same three lines identified above (Translation File Format).

A reference section will begin with a **<refset>** tag which contains a set of documents. Each document, defined by the **<doc>** tag, contains a set of segments. Each segment, defined by the **<seg>** tag, contains the translated text.

The <**refset>** tag has two required attributes "**setid**" and "**srclang**", and one implied attribute "**trglang"**. The **setid** attribute contains the name of the document set that has been translated. This name will match the **setid** of the source file for which human translators performed the translation. The **srclang** attribute indicates the language of the source set and the **trglang** attribute indicates the language of the translate set, for Metric MATR this will be set to "**English"**.

The **<doc>** tag has two required attributes "**docid**" and "**genre**", and one implied attribute "**sysid**". The **docid** attribute contains the name identifying the document within the given source set. The genre attribute indicates the type of data for a given document. The **sysid** attribute contains the name of the human translator that performed the translation.

The **<seg>** tag has an implied attribute called **id**. The **id** attribute contains a number identifying the segment within the given document

4. <refset setid="mm08_set1_v0" srclang="Arabic" trglang="English">
5. <doc docid="document-1" genre="newswire" sysid="mm08_set1_system1">
6. <seg id="1"> TRANSLATED ENGLISH TEXT. </seg>
7. <seg id="2"> TRANSLATED ENGLISH TEXT. </seg>
8. …
9. </doc>
10. <doc docid="document-2" genre="newswire" sysid="mm08_set1_systems">
11. …
12. </doc>
13. …

14. </tstset>
15. </mteval>

Line 4: TSTSET tag identifies the beginning of a document list
Line 5: DOC tag, one for each document in the TSTSET
Line 6: SEG tag, ordered 1-N, for each sentence like unit in the document
Line 8: Many possible segments per document
Line 9: Closing DOC tag ends translations for the particular document
Line 13: Many possible documents in each TSTSET
Line 14: Closing TSTSET
Line 15: Closing MTEVAL tag

III. Source File Format

It is not anticipated that the developed metrics will require the use of the SOURCE transcripts. The XML file format description is included here for completeness.